Approximating Inference on Complex Motion Models Using Multi-model Particle Filter

Jianyu Wang[1], Debin Zhao[1,2], Shiguang Shan[1], and Wen Gao[1,2]

[1] Department of Computer Science, Harbin Institute of Technology, China
[2] JDL, Institute of Computing Technology, China Academy of Sciences
{jywang,dbzhao,sgshan,wgao}@jdl.ac.cn

Abstract. Due to its great ability of conquering clutters, which is especially useful for high-dimensional tracking problems, particle filter becomes popular in the visual tracking community. One remained difficulty of applying the particle filter to high-dimensional tracking problems is how to propagate particles efficiently considering complex motions of the target. In this paper, we propose the idea of approximating the complex motion model using a set of simple motion models to deal with the tracking problems cumbered by complex motions. Then, we provide a practical way to do inference on the set of simple motion models instead of original complex motion model in the particle filter. This new variation of particle filter is termed as Multi-Model Particle Filter (MMPF). We apply our proposed MMPF to the problem of head motion tracking. Note that the defined head motions include both rigid motions and non-rigid motions. Experiments show that, when compared with the standard particle filter, the MMPF works well for this high-dimensional tracking problem with reasonable computational cost. In addition, the MMPF may provide a possible solution to other high-dimensional sequential state estimation problems such as human body pose estimation and sign language estimation and recognition from video.

1 Introduction

Many researchers make extensive efforts in the visual tracking area and two decades of research have yielded many powerful tracking systems [1,3,4,5,6,8, e.g.]. One remained challenging problem for visual tracking is how to deal with tracking problems cumbered by high-dimensional complex motions robustly and efficiently, such as non-rigid head motion estimation, hand pose and body pose recognition.

Due to doing inference under the Bayesian framework and not assuming the distribution form of the posterior, particle filter, also known as CONDENSION in the computer vision community [1,4], becomes popular and is one of the promising techniques to deal with complex tracking problems with the ability of integrating different cues of information. When applying particle filter to a specific task, one key component need to be carefully defined is dynamic models, which characterize the motion of the target and determines how the particles are

K. Aizawa, Y. Nakamura, and S. Satoh (Eds.): PCM 2004, LNCS 3332, pp. 1011–1018, 2004.

propagated in the state space. Only when the particles are properly propagated, satisfied posterior may be obtained sequentially.

Previous works either choose complex dynamic models or simple dynamic models to characterize the target dynamics and to propagate particles [4,6,8,9]. The advantages of simple models are that they can be easily obtained and adapted to a specific application. Compared with complex models, simple models often show more elastic and robust with respect to noise since the states that can be reached are not carved tightly. Nevertheless, for high-dimensional tracking problems, simple models result in the most of particles with low weights and the efficiency of computation is low. As the dimension goes high, the exponential increasing computational burden quickly becomes prohibitly high to prevent simple models into practical use.

On the other hand, complex dynamic models incorporate more specific knowledge of how the object behaves than simple dynamic models. Therefore, it is more suitable for the high-dimensional tracking problems since computational cost is the key factor. Nevertheless, the complex dynamic models are typically learned from training examples or handcrafted using empirical knowledge. They are therefore very specific to the given task and are not easy to be obtained and adapted. Therefore, complex dynamic models are often learned by restricting the range of movement of the object and are easy to violate from the truth, e.g. assuming only walking or cycle motion can be handled for human body pose estimation [10]. These restrictions greatly reduce the generality of the resulting trackers.

In this paper, we propose a practical way to approaching the high-dimensional tracking problems which cumbered by complex motions. The main contributions of this paper can be concluded as follows:

First, we propose to using a set of simple motion models to approximating original strong motion models to ease the high dimensional curse.;

Then, we provide a practical solution of how to do inference by integrating multiple simple models in the particle filter.

Finally, we apply our proposed MMPF to the head motion tracking application. The experimental results show that the MMPF works well to this high-dimensional head motion tracking problem with reasonable computational resource.

The rest paper is organized as follows: In Section 2, We propose to approximate complex motion models using a set of simple motion models and to do integrated inference under the particle filter framework. We give the experiment results in Section 3 and conclude our work in Section 4.

2 Multi-model Particle Filter

Particle Filter is a technique for implementing a recursive temporal Bayesian filter by Monte Carlo simulations. The key idea is to represent the required posterior by a set of random samples and their associated weights. As the number of samples becomes sufficiently large, this Monte Carlo characterization becomes

an equivalent representation to the usual functional description of the posterior, and the particle filter approaches the optimal Bayesian estimate.

The power of the particle filter is in that it maintains a pool of hypotheses by sampling the proposal distribution $P(x_{i+1}|x_i)$ under the Bayesian framework. Generally, the more the hypotheses, the more chances to get accurate tracking results but the more computational resource is required. As the state of the object goes high, the computational cost quickly becomes prohibitively heavy due to the exponential computational complexity.

Following analysis can make this problem clear. To evaluate the efficiency of some particle set $\{x_i, \pi_i | i = 1, ..., n\}$, two measurements are defined [4]. One is the survival diagnostic $D = (\sum_{i=1}^{n} \pi_i^2)^{-1}$, another is the survival rate $\alpha \approx \frac{D}{n}$, where n is the number of particles. To guarantee the performance, it can be inferred that the required number n of particles should be $n \geq \frac{D_{min}}{\alpha^d}$, where D_{min} is the minimum acceptable survival diagnostic considering performance. It's clear that α^d is the determining factor of required particle number n and where the computational difficulties mainly arise from the dimension d. Therefore, directly apply the particle filter to high-dimensional tracking problem is computational intractable.

The particle filter's property of generating a set of hypotheses provides a natural way to approximating complex motion model using a set of simple motion models to generate several kinds of hypotheses in the pool instead of only one kind: instead of propagate all particles using the original complex motion model, the particle set are branched and each sub set of particles are propagated using one simple motion model. The final result is obtained by composite all the estimates using graphical model probabilistically.

In the following paragraph, the proposed MMPF is described in detail mathematically. We first define the following terms:

P_{mn}^{i-} : The probability at time i that the complex motion will be explained from simple model m to simple model n due to variation of target dynamics. These probabilities are assumed to be known in prior here and satisfy $\sum_{m=1}^{M} P_{mn}^{i-} = 1$, where M is the number of simple models. A state transition matrix M_{mat}^{i-}, which stacks the P_{mn}^{i-}, combines M simple models according to a graphic model under the Markov assumption:

$$M_{mat}^{i-} = \begin{bmatrix} P_{11}^{i-} & ... & P_{1M}^{i-} \\ ... & ... & ... \\ P_{M1}^{i-} & ... & P_{MM}^{i-} \end{bmatrix}. \tag{1}$$

P_{mn}^{i+}: The conditional probability that the target dynamics was explained from simple motion model m to simple motion model n at time i. Previous two probabilities describe how the simple models interact with each other to explain the complex motion model together.

P_{m}^{i-} : The probability that the target's dynamic will be explained by simple model m during time interval $[i, i+1)$ and satisfy $\sum_{m=1}^{M} P_{m}^{i-} = 1$.

P_{m}^{i+}: The probability after simple models' interaction that the target dynamics can be explained by simple model m and satisfy $\sum_{m=1}^{M} P_{m}^{i+} = 1$.

Let $s_i = \{x_i^k, w_i^k, m_i^k | k = 1, ..., N\}$ denote a particle set at time i, where m_i^k means the simple model according to which the particle k evolves in the state space at time i. For each particle, we define its private dynamic model according to the model probability P_m^{i-} and approximately there have the relation that the number of particles that will translate according to the simple model m is proportional to P_m^{i-}. That means all particles are divided into M groups probabilistically. Then, each group of particles behaves like a standard particle filter and M filtered states are obtained. Then the MMPF does an interaction between all filtered estimates and gets the final output by weighting all estimates statistically. After that, the model probability is updated according to the statistical property of residual error. The distance d_{i+1}^m which measures the residual error is application dependent and the distance we adopt to solve face tracking problem can be found in section 3.

Details of the algorithm are shown in Figure 1.

3 Application to Head Motion Tracking

In this section, we apply the MMPF method to head motion tracking with both rigid and non-rigid motions considered. Two difficulties are anticipated to be well handled under such a framework: one is that the method can work well with low quality image sequences and the other is that the tracker hold a high probability to recover from drift without manual re-initialization.

Experiments are performed on the real videos to test the tracker's ability of conquering clutters. The difficulties of tasks lie in that the states of the head need to be tracked are as high as 66 dimensions, which make the task is very challenging. While using MMPF, the original high-dimensional motions are factorized into eight simple models and make the problem tractable.

The experimental results show that the merits of this method can be concluded as follows:

1). The MMPF can be deal with low quality images due to the top-down matching scheme and stochastic search scheme (note that the experimental data we use are recorded using common hand held cameras and it was not high quality).

2). It is very robust to clutter. Even several frames are not well estimated and drift happen, the tracker holds a high probability to recover from the error. This is the essential merit for long sequence tracking in heavy clutter.

3.1 Face State Representation

A MPEG-4 compatible 3D parametric head model is implemented for synthesizing photo-realistic facial animations. One set of parameters can totally control the head motion and facial animations, named as Facial Animation Parameter (FAP) [11]. Here 66 low level FAPs are adopted instead of all 68 FAPs. In the experiments, we have made a try to use the CONDENSATION to do inference

on the 66 dimensional state space directly and stabilized results are not obtained even 10^8 particles are employed. The main reason is that most particles are wasted to generate useless hypotheses due to poor guidance. [2] has pointed out that when in spaces of dimensions much greater than about 10, good results are extremely difficult to get.

Iterate

Prediction: Sample N_m particles $s_i = \{x_i^k, w_i^k, m_i^k = m | k = 1, ..., N_m\}$ from simple model $P_m(x_t|x_{t-1})$, satisfying that $\frac{N_m}{N}$ is proportional to P_m^{i-} and $\sum_{m=1}^{M} N_m = N$.

Verification: Evaluate weights of N_m particles according to the likelihood model $w_t^k = P_m(y_t|x_t^{k,-})w_{t-1}^k$.

Interaction: 1). Compute an estimated state $\widehat{x}_t^m = \sum_{j=1}^{N_m} \frac{w_j^k x_j^k}{w^k}$ for sub-model $P_m(x_t|x_{t-1})$. 2). Compute the model probability $P_m^{t+} = \sum_{m=1}^{M} P_{mn}^{t-} P_m^{t-}$. 3). Compute the particle translation probability $P_{mn}^{t+} = P_{mn}^{t-} P_m^{t-} / P_m^{t+}$. 4). Compute M filtered estimated states $\widetilde{x}_{t+1}^n = \sum_{m=1}^{M} P_{mn}^{(t+1)+} \widehat{x}_{t+1}^m$. 5). Then the final result is $x_{t+1} = \sum_{n=1}^{M} P_n^{(t+1)+} \widetilde{x}_{t+1}^n$.

Updating Mode Probability: Compute distance d_t^m according to $P_m(z_t|x_t)$for each model m and update its probability $P_m^{(t+1)-} = V_{t+1}^m P_m^{t+}/C$, where $V_{t+1}^m = \frac{\exp(-(d_{t+1}^m)^2/2)}{\sqrt{(2\pi)^R \sigma_{t+1}^m}}$ and C is a normalizing constant.

Re-sampling: Compute the covariance of the normalized weights. If this variance exceeds some threshold, then construct a new set of samples by drawing, with replacement, N samples from the old set, using the weights as the probability that a sample will be drawn. The weight of each sample is now $\frac{1}{N}$.

Fig. 1. The process of Multi-Model Particle Filter

Since previous experiment denied the particle filter with the original high-dimensional state representation as a practical solution. We first do a dimension reduction using PCA to get intrinsic representations of the face motion state with that the head pose parameters are canceled out by setting them to zero. The first five eigen-values in descending order are retained to accommodate 99% variation of the training data set. (The training data are obtained by manually turned). The head pose dynamics are modeled using three Nearly Constant Velocity Models (NCVM). Therefore, the final state is the coefficients of the

five eigen-vectors, which characterize facial expressions and three coefficients of NCVM models, span an eight dimensional sub-space. We combine two kinds of simple dynamic models to approximate the original complex motion models in the MMPF algorithm.

3.2 The Set of Simple Dynamic Models

In previous sub-section, five eigen-vectors obtained by dimension reduction technique and three NCVMs which corresponding to head yaw, tilt and roll respectively are chosen as sub models in the MMPF. Consequently, the original unknown motion model is factorized into eight simple models and each simple model varies only in one dimension,

$$P(x_i|x_{i-1}) = \sum_{m=1}^{8} w_m P_m(x_i|x_{i-1}) \tag{2}$$

where $P_m(x_i|x_{i-1})$ represents one simple model and $P(x_i|x_{i-1})$ is the original strong motion model.

One step of the estimation process in experiments can be roughly represented by figure 2. The simple models are chained to do estimation in a cascade manner and construct a degenerate case of the MMPF. The previous estimated weight w_{m-1} of one simple model provides a starting point for the weight w_m's estimation of next sub-model on the chain.

The model interaction matrix is assumed to be constant in the estimation process and set as that the diagonal elements of the matrix are 0.72 and the non-diagonal elements are set to 0.04 (In our experiments, the results are not sensitive to the small variation of these parameters). The prior probability P_m^- is initially set to $\frac{1}{M}$. In the tracking process, the distance measuring residual error of simple model m in frame $i+1$ is set to

$$d_{i+1}^m = \frac{1}{|w_m^i - w_m^{i-1}|} \tag{3}$$

where w_m^i is the estimated weight of the simple model m in frame i.

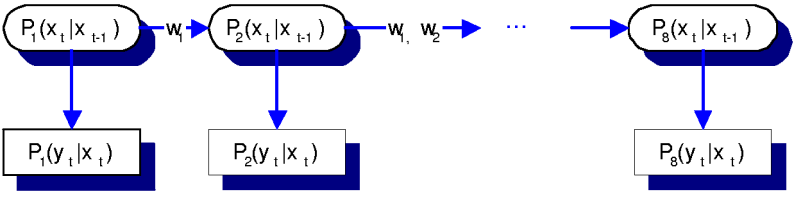

Fig. 2. The inference structure used for face tracking problem

3.3 Evaluating the Particles' Weights

In this sub-section, the likelihood model is constructed to relate the face states and face images and particles' weights are evaluated. When a frame I_t comes, the different image $\triangle I_t = I_t - I_{t-1}$ is first computed with the non-face area segmented out, where the I_{t-1} is the previous frame (all images are aligned manually according to the key feature points). Furthermore, $\triangle I_t$ is normalized like a matched filter to satisfy that: $||\triangle I_t|| = 0$ and $var(\triangle I_t) = 1$. Then, the tracker propagates particles to generate a pool of hypotheses. For each new hypothesis h_t^i generated by particle k_i, an observation image O_t^i is generated by the previous mentioned 3D face model. Also, a difference image $\triangle O_t^i$ between the O_t^i and I_{t-1} is computed considering the face area for each O_t^i. Then a dot product between $\triangle O_t^i$ and $\triangle I_t$ is calculated as the corresponding particle's weight

$$w_t^i = \triangle I_t \cdot \triangle O_t^i = \langle \triangle I_t, \triangle O_t^i \rangle \tag{4}$$

3.4 Qualitative Performance Evaluation Using Real Video Data

Four video footages corresponding to four persons' facial animation are recorded to test our algorithm's ability to conquer clutter and the performance under real world conditions. For the limit of space, only one video is shown in figure.3. It has 189 frames and is at the resolution of 320X240. Note that there are some difference between the reconstructed 3D face model and the person himself due to the reconstruction error. The stabilized results are obtained by employing 6800 particles. The top row of figure 3 shows the sample frames of the recorded video and the bottom row shows the corresponding re-synthesizing frames by estimated parameters.

To test the tracker's ability to conquer clutter, we also disturb the tracker's estimates with the noise during tracking. Averagely, for each coefficients, 18% of the 3σ violation from the right value parameters value, the tracker can quickly recover from the drift within three frames with the probability of 90.5% during 500 times tests, where the σ is the standard variance learned from the training set.

Fig. 3. Comparing original frames with re-synthesized frames

4 Conclusions and Future Work

In this paper, we propose a novel method to deal with the tracking problem suffered from high-dimensional complex motions. Do inference in Bayesian filter frame-work, more information can be incorporated under this framework to promote the performance of the tracker.

Future works includes:

1) Composing low level cues, such as optical flow or other motion estimation techniques to guiding how to propagate particles and thus accelerate the running speed of the system to achieve near real-time performance;

2) Using 3D facial morphable model [12] to automatic initialization of the head motion tracking system.

Acknowledgements. This research is partially supported by National Hi-Tech Program of China (No.2001AA114190 and No. 2002AA118010), National Nature Science Foundation of China (No. 60332010), and ISVISION Technologies Co. Ltd.

References

1. M. Isard and A. Blake, CONDENSATION – conditional density propagation for visual tracking, In Internal Journal of Computer Vision, 29, 1, 5–28, 1998.
2. David A. Forsyth, Jean Ponce, Computer Vision: A modern approach, published by Prentice Hall, 2002.
3. D. Comaniciu, V. Ramesh and P. Meer, Real-time tracking of non-rigid objects using Mean Shift, In IEEE Proceedings of CVPR, Hilton Head Island, South Carolina, Vol. 2, 142-149, 2000.
4. J. MacCormick and M. Isard, Partitioned sampling, articulated objects, and interface-quality hand tracker, In ECCV, Vol.2, pp.3-19, 2000.
5. C. Rasmussen and G. Hager, Probabilistic Data Association Methods for Tracking Complex Visual Objects, IEEE Transactions on PAMI, Vol. 23, No. 6, June 2001.
6. J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealled particle filtering. In IEEE Proceedings of CVPR, Hilton Head, V II pp. 126-133, 2000.
7. Kiam Choo and David J. Fleet. People tracking using hybrid monte carlo filtering. In IJCV, 2001.
8. H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In ECCV, Vol.2, pages 702-718, 2000.
9. Vladimir Pavlovic, James M. Rehg, Tat-Jen Cham, and Kevin P. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In ICCV,1999.
10. Rohr, K. Human movement analysis based on explicit motion models, In Motion-Based Recognition, kluwer Aademic Publishers, Dordrecht Boston, 1997, ch.8, 171-198.
11. J.Ostermann, Animation of Synthetic Faces in MPEG-4, Computer Animation, pp.49-51, June 8-10, 1998.
12. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces", In Proc. of SIGGRAPH99, 1999.